Parametric Joint Confidence Level Analysis: A Practical Cost and Schedule Risk Management Approach

Sara Jardine

Christian B. Smart, Ph.D., CCEA

Kimberly Roye, CCEA

Abstract: The use of Joint Confidence Level (JCL) analysis at NASA has proven to be a successful policy. Bottom-up resource-loaded schedules are the most common method for jointly analyzing cost and schedule risk. However, high-level parametrics and machine learning for JCL have been used successfully by one of the authors. This approach has some advantages over the more detailed method. In this paper, we discuss the use of parametrics and machine learning methods, especially as they apply to JCL analysis. The parametric and machine learning approach involves the development of mathematical models for cost and schedule risk. Parametric methods for cost typically use linear and nonlinear regression analysis. These methods applied to schedule often do not provide the high R-squared values seen in cost models. We discuss the application of machine learning models, such as regression trees, to develop higher-fidelity schedule models. We then introduce a bivariate model to combine the results of the cost and schedule risk analyses, along with correlation, to create a JCL using models for cost and schedule as inputs. We provide a previous case study of the successful use of this approach for a completed spacecraft mission and apply the approach to a large data set of cost, schedule, and technical information for software projects.

Background

For over fifty years, the cost analysis community has applied uncertainty analysis methods using univariate probability theory in risk analysis to generate separate distributions of a program's estimated cost and schedule (*Garvey*, 2000). In the schedule analysis and broader project management professional communities, the use of the schedule risk analysis has also been around for even longer and dates back to the Project Evaluation and Review Technique (*Hulett*, 2009). The interdependency between cost and schedule has long been recognized, but NASA is one of the few government agencies that has established official policy to conduct integrated cost and schedule risk analysis, which they call "joint confidence level analysis." We will use the term **joint confidence level** and its common abbreviation **JCL** throughout this paper.

The use of joint cost and schedule risk analysis has largely been limited to resource-loaded schedule analyses. While providing a great deal of insight into a project, resource-loaded schedules are labor-intensive. They also suffer from a drawback common to most bottom-up methods, which is the underestimation of the true amount of cost and schedule risk for a program. Parametric models can be developed much quicker and can provide a more comprehensive picture of program risk. Despite the development of such methods more than 20 years ago (*Garvey*, 2000), little has been adopted from multivariate theory to combine or develop conditional cost and schedule probability distributions to present to decision-makers.

Introduction

This paper reintroduces the top-down parametric approach to conducting JCL analysis. This technique is less cumbersome yet just as accurate

in the quantification of top-level cost and schedule risk as the familiar bottom-up resource loaded JCL method. We enhance the practice of the top-down parametric method with the consideration of machine learning techniques in addition to the use of traditional parametric regression analysis. We introduce the application of optimization methods to develop Cost Estimating Relationships (CER). We present regression trees as a means to develop better Schedule Estimating Relationships (SER), since it is more difficult to use traditional regression methods to derive meaningful trendlines using historical schedule data. Using the results of the individual cost and schedule analysis, uncertainty analysis is applied separately to compute the means and variances, which are used to specify the parameters of a bivariate probability model for a given program. Dr. Christian Smart has developed a standalone MS Excel spreadsheet to compute a bivariate probability model. Using the means and variances from the Cost Risk Analysis (CRA) and Schedule Risk Analysis (SRA) along with the program's target budget and schedule values, the calculator will produce the JCL and associated iso-curves at various joint confidence levels.

In this paper, the following topics are addressed:

- Benefits of JCL within Project Management
- JCL Methods: Bottoms-Up and Top-Down Parametric
- Parametric Machine Learning Techniques: Optimization and Regression Trees
- Top-Down Parametric Method Case Study: NASA MAVEN spacecraft program

In summary, this paper highlights the benefits of JCL analysis and offers a quicker top-down parametric JCL method to be used by the cost community. The JCL provides a more holistic view of uncertainty so that decision-makers can make more informed decisions. We provide a comparison of the top-down and more well-known bottom-up JCL approaches, provide an in-depth process for the top-down JCL method using a software program example, and demonstrate a real-life successful NASA spacecraft program that used the top-down parametric JCL approach.

Joint Confidence Level Benefits to Risk Management

Projects of all types frequently experience cost growth and schedule delays. Projects that do not suffer from one or both maladies are the rare exception, rather than rule. In addition to being common, these phenomena are often extreme, especially for cost. Indeed, the cost for approximately 1 in 6 defense and NASA missions doubles or more from the initial plan to the final actual. Defense and NASA projects are comparable to other industries, as shown in Table 1. These issues are long-standing and have shown no signs of improving over the last several decades.

	Olympics	Software/ IT	Dams	NASA/ DoD	Rail	Bridges/ Tunnels	Roads
Average Cost Growth	156%	43–56%	24–96%	52%	45%	34%	20%
Frequency of Occurrence	10/10	8/10	8/10	8/10	9/10	9/10	9/10
Frequency of Doubling	1 in 2	1 in 4	1 in 5	1 in 6	1 in 12	1 in 12	1 in 50
Average Schedule Delay	0%	63–84%	27–44%	27–52%	45%	23%	38%
Frequency of Schedule Delay	0/10	9/10	7/10	9/10	8/10	7/10	7/10

Table 1. Comparison of Cost Growth and Schedule Delays Across Several Industries. (Source: Solving for Project Risk Management, Christian Smart, McGraw-Hill, 2020).

The extent and the frequency of cost increases and schedule slips is prima facie evidence that these programs have a significant amount of resource risk and that this risk has not been managed well. The resource risks for these projects have also not been analyzed with accuracy, as exhibited by the track record for cost and schedule risk analysis. For cost analysis, see Table 2 for a comparison of the 90% confidence levels (90th percentile of the CDF or S-curve) with the actual costs.

Project	Cost Growth	Ratio of Actual Cost to 90% Confidence Level
1	0%	0.6
2	19%	1.1
3	31%	1.0
4	32%	1.1
5	greater than 45%	greater than 1.0
6	52%	1.5
7	84%	1.7
8	93%	1.6
9	121%	2.0
10	280%	2.2

Table 2. Cost Growth and Ratio of Actual Cost to 90% Confidence Level for 10 Historical Projects (Source: Solving for Project Risk Management, Christian Smart, McGraw-Hill, 2020).

The projects in Table 2 are from a variety of applications. JCLs were conducted for at least two of the missions. For 5 of the 10 missions, the actual cost was at least one and a half times the 90% confidence level, and for 2 it was double or more. Two of the missions listed in the table were cancelled. If they had not been cancelled, the cost growth would have been higher. The term "90% confidence level" for these analyses is grossly erroneous. Even so, 90% confidence levels should have been high enough to capture these variations. However, the actual cost was greater than the 90% confidence level for 8 of the 10 projects. This dismal result is even worse than it appears. A more in depth discussion for projects one and five is provided below.

- **Project 1.** One of the authors conducted a cost and schedule risk analyses using the top-down parametric method for project 1, which was a relatively rare mission that did not experience cost growth. The estimate of the 50% confidence level was within 1% of the actual cost. The project also completed on time, in line with the 50% confidence level for schedule. This kind of outcome is the exception rather than the rule. As can be seen from the table, all the other missions experienced significant cost growth. This provides evidence that the parametric JCL approach may be better at capturing the full extent of resource risk.
- **Project 5.** This project experienced such significant growth from one phase to the next that it exceeded the 90% confidence level well before completion.

The National Aeronautics and Space Administration (NASA) is one of the few government agencies that requires a JCL analysis be conducted for programs and projects. A JCL analysis is a process that combines a program or project's cost, schedule, and risk into an integrated picture. It represents the probability that a program cost will be equal to or less than the targeted cost, and that the schedule will be equal to or less than the targeted finish date. According to the most recent NASA JCL policy, by providing a confidence level that integrates cost and schedule, the JCL helps inform management of the likelihood of a program's programmatic success. Implementing JCL requirements for NASA programs has proven to be an effective forcing function to help program managers integrate stove-piped work products such as an Integrated Master Schedule (IMS), resource management, and risk management (*NASA JCL Requirements Update Memo*, 2019).

A program manager's decision space encompasses cost, schedule, and performance of a program. Risk analysis is needed when the expectations in any of these domains limit what is feasible. Therefore, managing risk is to manage the conflicts that exists within each domain and interdependencies across all three (*Garvey*, 1993). Generating a joint probability distribution supports the estimation of a program's cost and schedule, which simultaneously have a specified probability of not being exceeded. Because it is a more stringent requirement, the JCL is almost always higher than either the cost or schedule confidence level when developed separately. The JCL provides program managers with an assessment of the likelihood of achieving a budget for a given schedule, which aids the creation and management of credible project plans. Depending on the agency's JCL goal, the amount of cost reserves and additional schedule

can be determined and provided to decision-makers. Project management can then more effectively manage scope, cost reserves and schedule reserves of the project to mitigate the risk.

Joint Confidence Level Methods

There are two proven processes to calculate a JCL: the bottom-up resource-loaded schedule method and the top-down parametric method. Although the intention of this paper is to encourage the use of the top-down parametric as a more practical approach in the cost estimating field, we will briefly discuss the bottom-up method for the purpose of comparing it to the top-down method.

Bottom-Up Method

The bottom-up JCL method starts with a robust cost estimate and is mapped to a resource-loaded Integrated Master Schedule (IMS). A risk list is incorporated in the joint cost and schedule model at the lowest WBS element level and schedule and cost uncertainty is assigned. Although the bottom-up method is popular and can successfully calculate a JCL, it has its disadvantages.

Shortcomings of the bottom-up JCL approach include being resource intensive and time-consuming. As with any bottom-up estimating approach, it is easier to inadvertently miss the accounting for uncertainty of lower-level risk elements and thus, underestimate risk of the overall program. It is also difficult to justify uncertainty probability distributions on lower-level elements since data is scarcer and is typically not available at a low level. The bottom-up method also ignores unknown-unknowns, which are largely covered in the historical parametric data used in the top-down approach. While unknown-unknowns cannot be predicted in advanced, their existence in the aggregate can be used in the quantification of cost and schedule risk with just as much confidence as actuaries place in the quantification of insurance risk (Augustine 1983). While they are impossible to predict in advance, they dominate the bulk of cost and schedule risk, so their inclusion is imperative in conducting realistic risk assessments. The inclusion of unknown-unknowns is largely captured by the standard error and prediction intervals derived from the parametric cost and schedule equations.

The 2014 Joint Agency Cost Schedule Risk and Uncertainty Handbook (JA CSRUH) highlights the Fully Integrated Cost and Schedule Method (FICSM) as a bottom-up JCL approach. To provide a general understanding of the time-intensive bottom-up process, the FICSM approach is illustrated in Figure 1 below. This method can be applied using Joint Analysis of Cost and Schedule (JACS) in the ACEIT software suite and MS Project.

Top-Down Method

The top-down parametric JCL approach is less resource intensive than the bottom-up approach.

Figure 1. FICSM Process (Source: Joint Agency Cost Schedule Risk and Uncertainty Handbook 2014).

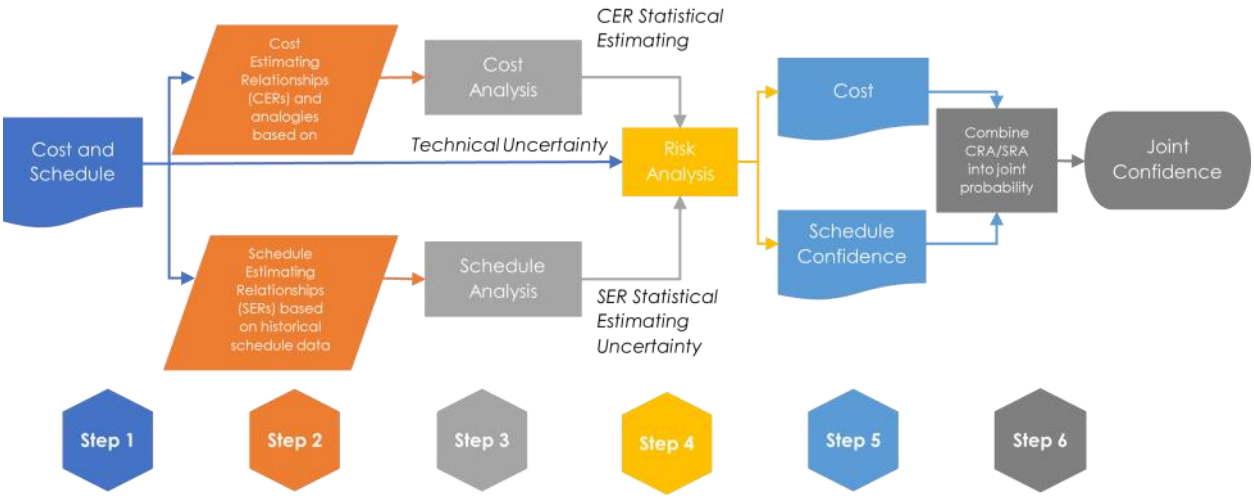

Figure 2. Top-Down Parametric JCL Process.

The reference to understand and explain the top-down parametric JCL approach was adopted from "*A Family of Joint Probability Models for Cost and Schedule Uncertainties*" (*Garvey*, 1993). To begin the discussion, an illustration of the top-down parametric process is illustrated in Figure 2.

A description for each of the six steps will be provided, while a more in-depth approach will be discussed for Step 2, where cost and schedule analyses are developed independently. During this step, if traditional parametric regression approaches do not result in any viable statistically significant estimating relationships, machine learning techniques can be used to predict estimating relationships. Throughout the steps, we will use a hypothetical software program example to demonstrate the top-down parametric JCL process.

Step 1: Cost and Schedule Data Collection. To begin, the analyst should collect a schedule and cost dataset separately that meets the criteria for performing parametric analysis to test the statistical significance of a cost and schedule estimating relationship. Data collection for the dataset would include historical analogous programs.

In the software program example, the cost dataset included hours as the dependent variable and peak staff and Equivalent Source Lines of Code (ESLOC) as the independent variables. The schedule dataset included duration in months as the dependent variable and potential schedule drivers such as new code, peak staff, and total development hours.

If data are not available, there are a variety of off-the-shelf parametric estimating tools that can be used including SEER-H, SEER-SEM, and SEER-Space.

Step 2: Cost and Schedule Regression Analysis. Perform regression analysis on the cost and schedule datasets separately using linear and nonlinear models. Test the statistical significance of regression equations and determine if any viable regression equations result. Different statistical software tools can be used to perform regression analysis during this step, including MS Excel, CO$TAT, or JMP. If traditional regression analysis does not result in any CERs or SERs, machine learning techniques should be considered.

Parametric techniques are within the scope of machine learning and can be applied to determine relationships between cost and

schedule and their drivers. These machine learning techniques include optimization to produce the "best" coefficients for a regression equation and regression trees. We introduce the discussion of regression trees in parametric estimating of schedules due to the fact that SERs are more difficult to estimate using traditional regression methods. The range of schedules typically has a smaller spread than cost, making trendlines less statistically significant. However, program technical data often includes a considerable amount of categorical data, which lends itself well to the use of regression trees. In a later section of this paper, we will provide a more in-depth discussion on the use optimization and regression trees for Step 2 of the top-down parametric approach.

In the software program example, optimization was applied using MS Excel Solver to develop a CER where peak staff and ESLOC were the independent variables driving hours. Since the example software program dataset was large (e.g., more than 50 data points), Maximum Likelihood Estimation Regression for Log Normal Error (MRLN or "Merlin") regression method was used (Smart 2017). MRLN will be further discussed in the next section to demonstrate how to apply optimization to determine the optimal coefficients, $\boldsymbol{\beta_0, \beta_1, \beta_2,}$ for the regression equation. With a Pearson's R^2 equal to 74%, the resulting CER had the following nonlinear power equation:

$$Total\ Hours = \boldsymbol{\beta_0}(Peak\ Staff)^{\beta_1}(ESLOC)^{\beta_2}$$

In the software program example, the schedule dataset did not result in any statistically significant SERs. With a significant amount of categorical data such as development process type (e.g., waterfall, incremental, agile, evolutionary, etc.), operating environment, and application domain, a regression tree with a Pearson's R^2 equal to 50% was developed using the *R* statistical programming platform.

Step 3: Cost and Schedule Analysis. This step represents the parametric results of the cost and schedule analyses developed in Step 2.

Step 4: Cost and Schedule Risk Analysis. Conduct a cost and schedule risk analysis on the cost and schedule estimate results, respectively. To achieve this step, a brief discussion of how to apply uncertainty analysis to regression equations is necessary. Regression equations have two forms of uncertainty that need to be accounted for: **input** and **estimating**.

Input uncertainty represents variability in the independent variables in a CER/SER regression equation. One approach to computing input uncertainty, *X*, is to assume a triangular distribution on input variables and run low (*L*), most likely (*ML*), and high (*H*) values through the CER/SER to obtain *L*, *ML*, and *H* estimates. Calculate the mean, $\boldsymbol{\mu_X}$, and standard deviation, $\boldsymbol{\sigma_X}$, of the triangular distribution. The calculations of the mean and standard deviation are:

$$\mu_X = \frac{L+ML+H}{3}$$
$$\sigma_X = \sqrt{\frac{L^2 + ML^2 + H^2 - L*ML - L*H - ML*H}{18}} \quad (1)$$

Estimating uncertainty is inherent to regression equations because, regardless of the parametric method used, even if the independent variables are known precisely, the CER/SER equation will return a result that is not certain. The error of the regression equation scales with the CER/SER result, making multiplicative error terms the preferred approach to modeling CER/SER estimating uncertainty. Regression estimating uncertainty represents uncertainty about the estimate's residual ε, (e.g., $Y = aX^b\varepsilon$). The farther the input variable is from the center of mass data used to derive the CER/SER, the greater the uncertainty of the CER/SER. The prediction interval or standard error provided by the

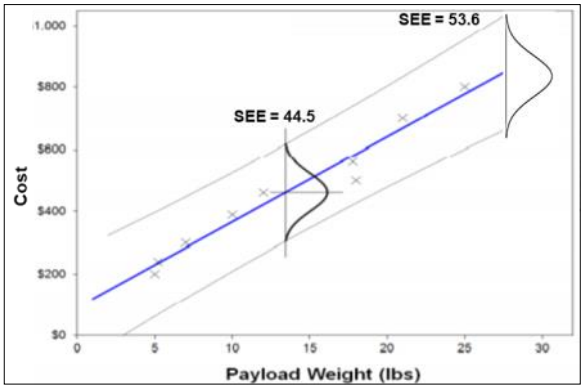

Figure 3. CER/SER Uncertainty Bounds (Source: Joint Agency Cost Schedule Risk and Uncertainty Handbook, 2014).

regression analysis can be used to determine the CER/SER uncertainty bounds. The Standard Error of the Estimate (SEE) converts to a prediction interval to account for the distance of the estimate from the center of the CER/SER dataset. Figure 3 shows a CER example of cost as a function of weight where uncertainty increases (standard deviation gets larger) as the point estimate moves towards the data boundaries (*JA CSRUH*, 2014).

One approach to computing estimating uncertainty, *Y*, is to treat uncertainty as a lognormal distribution and calculate the mean and standard deviation. Compute the mean and standard deviation in log space and then convert the values to unit space. The formulas to convert the mean, μ_y, and standard deviation, σ_y, from log to unit space are shown below:

$$\mu_y = e^{\mu+\frac{1}{2}\sigma^2}$$

$$\sigma_y = \sqrt{\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}} \qquad (2)$$

To compute the total uncertainty of the regression equation, input and estimating uncertainty can be combined using propagation of errors. Assuming input and estimating uncertainty are independent and the residuals are multiplicative, the total uncertainty is obtained by multiplying the means and standard deviations of the input and estimating uncertainty calculated in (1) and (2). The formulas to combine input, *X*, and estimating, *Y*, are shown below:

$$\mu(X*Y) = \mu_X * \mu_Y$$

$$\sigma(X*Y) = \sqrt{\sigma_x^2*\sigma_y^2 + \sigma_x^2\mu_y^2 + \sigma_y^2\mu_x^2} \qquad (3)$$

Figure 4 shows a graphical representation of combining input and estimating uncertainty of a CER/SER calculated in (3) referenced from the 2014 *JA CSRUH*.

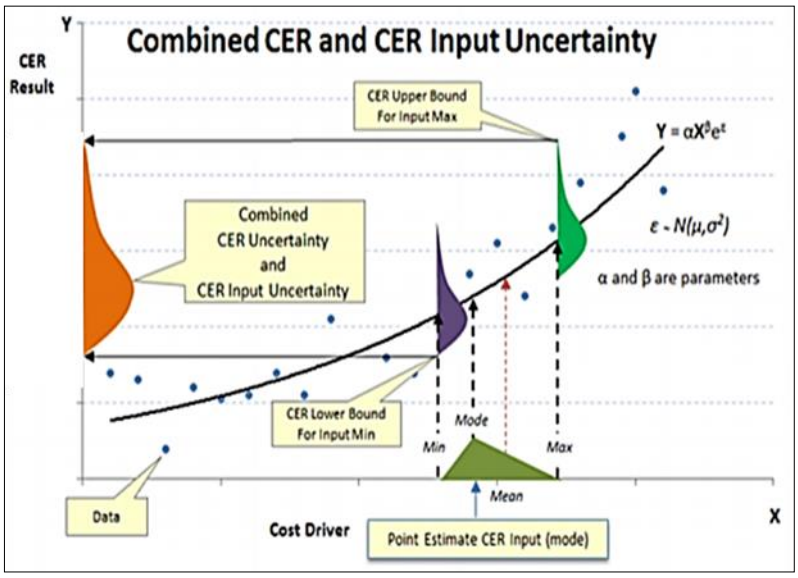

Figure 4. Combining Input and Estimating Uncertainty of a CER/SER. (Source: Joint Agency Cost Schedule Risk and Uncertainty Handbook, 2014).

If a regression tree was the method employed in Step 2 to derive a parametric relationship, as with the SER analysis done for the software program example, uncertainty analysis should be conducted on the regression tree. Input

uncertainty is modeled the same way as a regression equation using a triangular distribution on the input variables. To derive regression tree input and estimating uncertainty, assume they are independent. Depending on the data, the errors may be additive or multiplicative. In examining the data, we found the residuals best fit an additive model with a Gaussian/normal distribution. The input variables such as peak staff and software development hours are varied and simulated. For each trial, the simulation also sampled from the Gaussian for the regression tree residuals. For each of the 1,000 trials, the results from the varied input variables, and the estimation uncertainty from the residuals were added to yield total uncertainty.

Step 5: Cost and Schedule Confidence Levels. This step represents the results of the cost and schedule risk analyses (CRA/SRA) developed in Step 4. The results would reflect separate cumulative probability distributions or S-Curve results from the cost and schedule risk analyses.

Step: 6: Joint Confidence Level. The final step is to combine the CRA and SRA developed in Step 5 into a joint probability distribution to calculate the JCL. The reasons being are because they directly incorporate correlation between cost and schedule for programs and these distributions provide at least some probability of a cost or schedule overrun (lognormal distribution having a larger skew to the right while the normal distribution is not skewed). In accordance with Paul Garvey's method to combine cost and schedule as a joint probability model, we provide the following distributions to model the behavior of program cost and schedule: bivariate normal, bivariate lognormal, and bivariate normal-lognormal distributions. Figure 5 provides graphical depictions of a normal distribution.

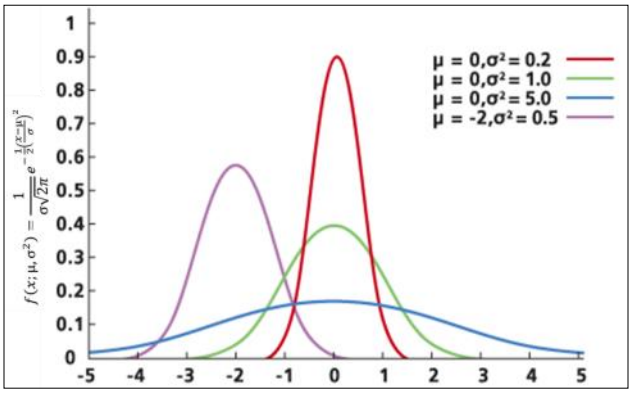

Figure 5. Gaussian/Normal Distribution.

In the authors' experience and results from the Air Force, Cost Risk Uncertainty Analysis Metrics Manual (CRUAMM), cost uncertainty is rarely normally distributed. When it comes to cost estimating the so-called normal is anything but normal! Cost estimating uncertainty is typically best modeled with a lognormal distribution. Schedule distribution uncertainty is typically lognormal, but in some instances, as in our example, the normal distribution is a good fit. The lognormal distribution is a skewed distribution. The lower bound is never less than zero meaning the cost and schedule cannot become negative and has an upper bound of infinity. The probability is skewed right providing at least some probability of a large cost or schedule overrun. These characteristics make the lognormal appealing for cost modeling and a best choice in the absence of better information (*JA CSRUH*, 2014). Figure 6 provides graphical depictions of lognormal distributions.

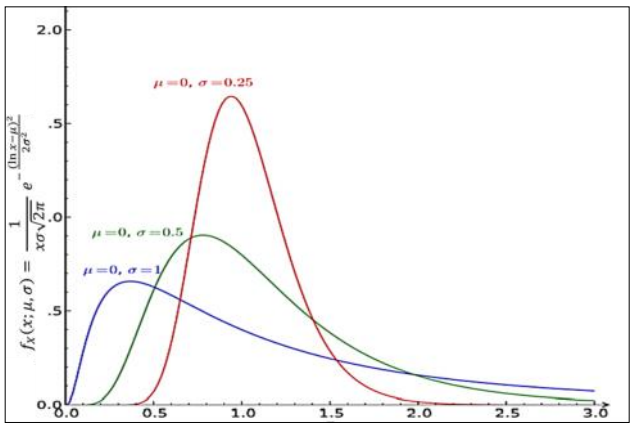

Figure 6. Lognormal Distribution.

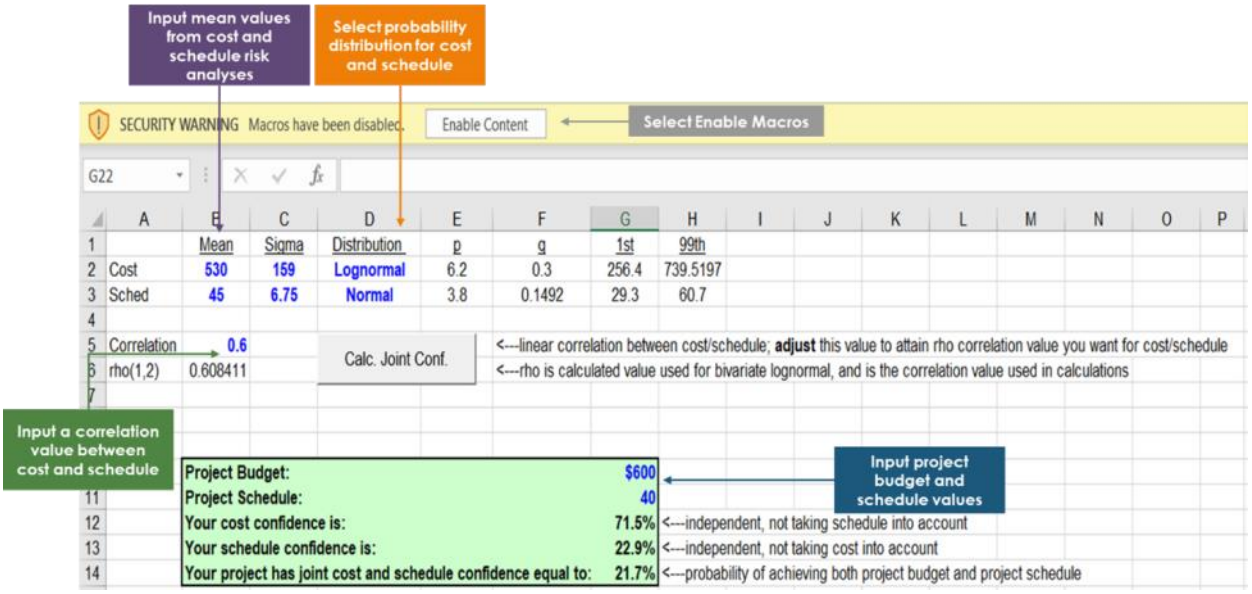

Figure 7. MS Excel JCL Calculator (Top-Down Parametric Method).

The bivariate normal-lognormal has two different marginal distributions. One marginal is normal, and the other is lognormal. Situations may arise when normal and lognormal distributions characterize a program's cost and schedule distributions.

To calculate a joint confidence, assume lognormal or normal risk distributions on cost and schedule using the mean and standard deviation as the parameters derived from the cost and schedule analyses in Step 5. Assume a positive linear correlation value between cost and schedule (e.g., correlation value equal to 0.6 or 0.7). Figure 7 is a screenshot of the MS Excel JCL calculator developed by one of the authors, Dr. Christian Smart, to take values derived from the top-down parametric approach and provide a JCL. The calculator uses a macro that numerically approximates the bivariate probability distribution, aka JCL, values.

In the notional example provided in Figure 7, the target budget of the given program is $600M and target schedule is 40 months. Using the results of the CRA and SRA, the mean and standard deviation is $530M and $159M respectively for cost, and 45 months and 6.75 months respectively for schedule. Cost is assumed to be lognormally distributed while schedule normally distributed. The correlation value between cost and schedule was selected to be 0.6. Based on the author's experience and data analysis, this is a reasonable value. The resulting JCL is 21.7%, meaning there is a 21.7% chance that the program cost will be equal to or less than $600M and that the schedule will be equal to or less than 40 months. If schedule was not

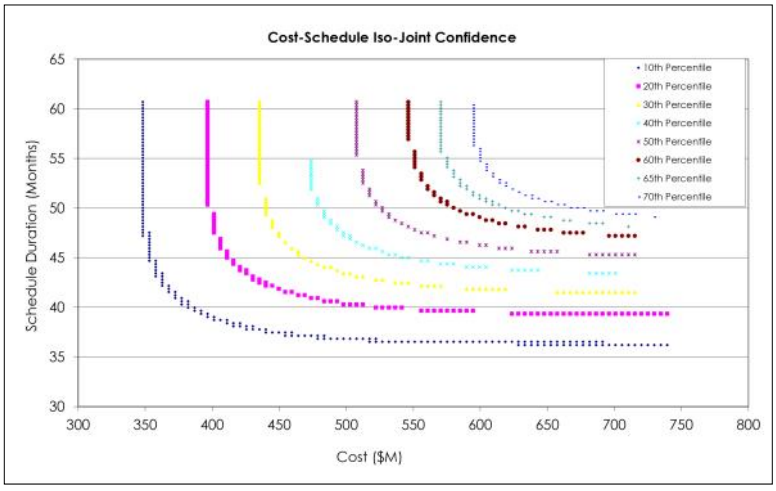

Figure 8. Example JCL Iso-Curves.

considered, the cost confidence level was 71.5% and if cost was not considered, the schedule confidence level was 22.9%.

Figure 8 shows the resulting JCL iso-curves with cost on the x-axis and schedule on the y-axis. As "iso" is a prefix meaning "equal", each cost and schedule joint iso-curve in the graph represents a specific JCL confidence level percentile. You can determine the current JCL level of your project by looking at the position of the budget and project planned schedule.

Figure 9 shows another example of a JCL output that would be presented to management.

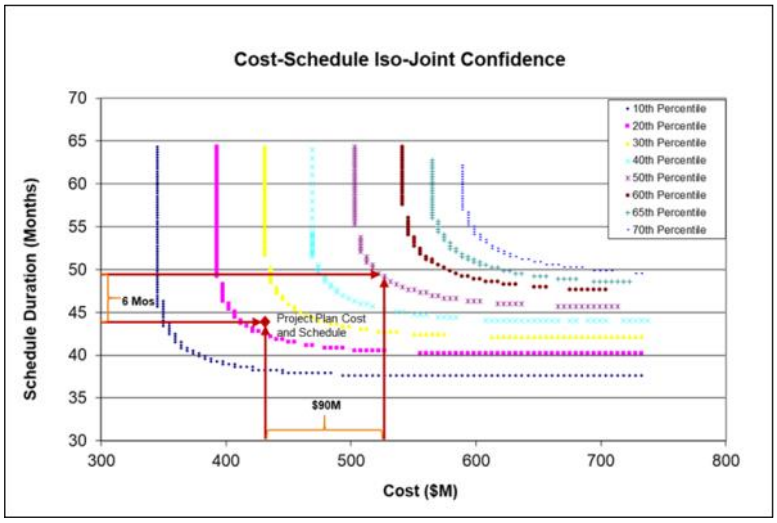

Figure 9. Example JCL Result.

The example program has a current project plan of a $435M budget and a 43-month schedule. Looking at the graph, this project plan position is at the 24% JCL. If management were interested in how much funding and schedule was needed to achieve the 50% JCL, you would look at the 50th percentile JCL iso-curve and see that an additional $90M and 6 months would be required. As there are multiple pairs of cost and schedule on each iso -curve, depending on the relative importance of schedule versus cost, an analyst can determine the amount of cost reserves and additional schedule duration needed to achieve the agency's JCL goal.

Machine Learning Techniques for Parametric Estimating

Machine learning methods can be a powerful mechanism to determine estimating relationships in a dataset when conducting the top-down parametric JCL. Machine learning is a collection of mathematical methods and computer algorithms for prediction and classification that represent a more modern way of conducting analysis on datasets that incorporates the use of computer programming with statistical analysis. Modern machine learning methods include decision trees, deep learning, and text analytics. As mentioned, machine learning techniques can be applied when developing a cost and schedule analysis during the top-down parametric JCL approach when traditional regression methods do not provide meaningful results, such as a regression equation with a low R^2, for example. For the purposes of this paper, we will focus on how to apply optimization and regression trees to develop cost and schedule estimating relationships when conducting a top-down parametric JCL analysis.

Optimization Technique for Parametric Estimating

Regression analysis as performed in Step 2 of the top-down JCL method, is a form of optimization. Optimization is a collection of mathematical principles and methods used for solving quantitative problems. The goal is to minimize or maximize a function in pursuit of finding the "best" solution. As previously mentioned, we will discuss the application of maximum likelihood as a regression approach to develop unbiased, optimal estimates of the mean when the errors are lognormally distributed. MRLN was developed by one of the authors, Dr. Christian Smart (*Smart*, 2017).

Let $a_1,\ldots, a_n$ represent the observed data and $x_1, \ldots,x_n$ represent random variables where a_i results from observing the random variable x_i. The likelihood function, which represents the likelihood of obtaining the sample data, is:

$$L(\theta) = \prod_{i=1}^{n} Pr(X_i = A_i|\theta)$$

The vector, θ, maximizes the likelihood function in the likelihood function. This consistent and efficient method is advantageous because maximizing the likelihood of finding the true underlying parameters of this distribution is exactly what we hope to accomplish in developing a CER. Other advantages of maximum likelihood are that it is always available to use, and it uses all the available data, where other methods such as percentile matching and method of moments do not.

Recall in the software program example, we estimated the following CER power equation model form:

$$Y = \beta_0 X_1^{\beta_1} \ldots X_p^{\beta_p}$$

The goal for MRLN is to maximize the function:

$$l(\beta_0, \beta_1, \ldots, \beta_p, \theta) = -\frac{n}{2} ln(\theta) - \frac{1}{2\,\theta} \sum_{i=1}^{n} \left(ln(y_i) - ln(\beta_0) - \sum_{j=1}^{p} \beta_j ln(X_{ij}) + \frac{\theta}{2} \right)^2$$

Using the MRLN method, MS Excel Solver can be used to find an optimal value in a cell. Decision variables are used to compute the formulas defined in the objective to converge on a solution that maximizes values for $\beta_0, \beta_1, \ldots, \beta_p$ to form the power equation, $Y = \beta_0 X_1^{\beta_1} \ldots X_p^{\beta_p}$. When using Excel Solver to optimize the coefficients in the software program using the MRLN regression method, recall that it resulted in the following CER with a Pearson's R^2 equal to 74%:

$$Total\ Hours = \boldsymbol{\beta_0}(Peak\ Staff)^{\beta_1}(ESLOC)^{\beta_2}$$

Regression Tree Technique for Parametric Estimating

Regression trees are an effective way to visualize the relationships between features within datasets, particularly when there is a large amount of categorical data such as historical schedule datasets. Regression trees can be used in preliminary data exploration to understand the most significant variables within a dataset. Regression trees can also be used to show the relationships within a dataset in Step 2 of the top-down JCL method when traditional regression analysis does not produce any good results. Pairwise analysis combined with regression trees can help shorten the time running regression models in search of significant relationships. Two of the authors, Kimberly Roye and Dr. Christian Smart, provided an overview of regression trees in a 2019 ICEAA presentation (*Roye and Smart,* 2019).

In a regression tree, the data are split into homogenous groups, and the graphs present splits with the use of branches (called decision nodes) and leaves (terminal nodes). The goal of a regression tree is to partition data into smaller regions where interactions are more manageable. They are useful when there is a non-linear and complex relationship between dependent and independent variables that cannot otherwise be represented by a regression equation. Figure 10 illustrates the structure of a regression tree.

The root node represents sample dataset that is being analyzed. The method asks its first yes or no question and splits the data into two groups based on the answer. The decision nodes represent the first set of homogenous groups discovered within the dataset. On the left, another yes or no question is asked, and the group splits into two nodes: one terminal and one decision node. The criterion for splitting is the choice that reduces the sum of squared errors by the biggest

amount. This process is recursively applied to each of the subsets produced until the reduction in error is smaller than a pre-specified limit, such as 1x10^-5. When a decision node be can split no further, the branch ends in a leaf, or terminal node. Each terminal node is a subset of the data set, and the estimate at each terminal node is the average of the data points in that subset.

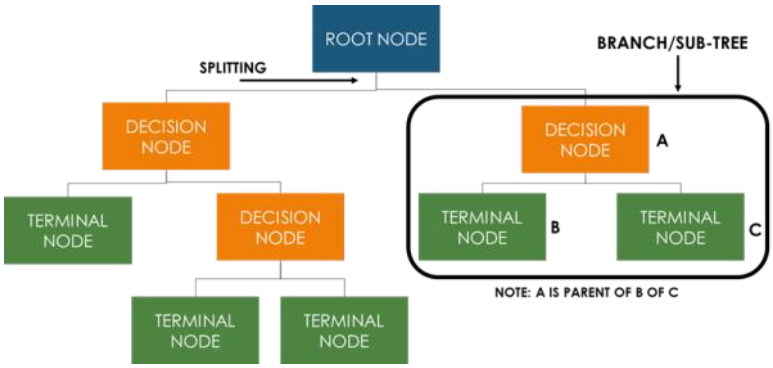

Figure 10. Regression Tree Layout.

In our software program example, since no SERs were significant in Step 2 of the top-down parametric approach, the schedule dataset was used to develop a regression tree using the *R* statistical programming platform. In the software program schedule dataset, total software development hours proved to be the most important factor. With a Pearson's R^2 approximately equal to 50%, Figure 11 shows an abbreviated version of the resulting schedule regression tree for the software program.

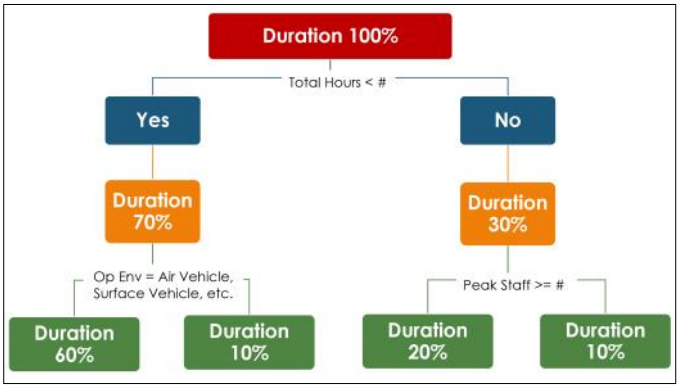

Figure 11. Software Program Example Regression Tree.

To explain this software program regression tree, we start with the total (100%) schedule dataset. Next, we ask if the total number of development hours is more than or less than a duration value. If the answer is yes, then the data is split into a branch to the left and if the answer is no, then the data is split into the branch on the right. In this example, 70% of the data satisfied the condition for number of hours and 30% did not. Total software development hours best minimize the squared error when estimating schedule duration. There are three decision nodes that ask questions about the value of total software development hours. Based on the value of total software development hours, we end at one of the terminal nodes of the tree. The estimate at the terminal, or leaf, node is the average duration of the subset included in that node. Each split is labeled with a condition and the branches between them are labeled with the average duration for that dataset split. An example interpretation of the first decision in the tree is, "if the total software development hours is more than 50,000 hours, my estimate is 60 months, otherwise it is 40 months." The regression tree produces a point estimate. Just like with traditional regression analysis, the regression tree uncertainty analysis is conducted by assessing the residuals, fitting a distribution, and combining this with parameter uncertainty, which provides an overall uncertainty distribution for the parametric schedule estimate.

For each node in the tree, the regression tree split is chosen by the algorithm to minimize the sum of squared errors. The algorithm chooses the variable and the associated value based on what reduces the sum of squared errors the most.

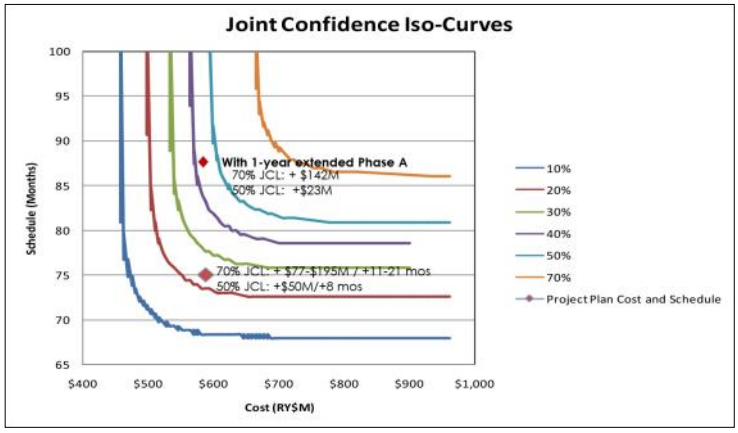

Figure 12. NASA Maven Spacecraft Program JCL Estimate.

Case Study

To advocate for and demonstrate the effectiveness of the top-down parametric approach, we highlight a success story for the NASA MAVEN spacecraft program. In 2009, one of the authors developed a JCL using the top-down parametric approach. At the time, with the project plan cost and schedule, the JCL was estimated at 23% and if a year was added to the development schedule, the JCL was estimated at 44%. With the current project plan, to achieve a JCL of 50%, an additional $50M and eight months would be needed, while to achieve a 70% JCL, an additional $77M-195M and 11-21 months would be needed. In 2013, the actuals for cost and schedule for the Maven program came in at the 50% JCL that was estimated in 2009. This is one of the few programs to show no cost growth, demonstrating an estimate that actually "hit the mark" when funded to the predicted 50% JCL. Table 3 summarizes the JCL results estimated in 2009 and the actual results in 2013.

	2009 (Estimate)			2013 (Actuals)
	Project Plan	50% JCL	70% JCL	
Cost ($M)	~$590	$640	$667-$785	$640
Schedule (months)	~75	83	86-96	83

Table 3. MAVEN Program JCL Results.

Figure 12 shows the iso-curves calculated for the MAVEN program.

MAVEN was the mission in Table 2 presented earlier for which the actual cost was below the 90th percentile of the cost risk analysis. While only one data point, it provides evidence that parametric JCLs can help ensure credible risk ranges.

Conclusion

NASA is the only known government agency that currently has a JCL policy. Very few organizations perform JCL analysis routinely as part of project management decision making. The more informed and holistic cost and schedule risk analysis results of a JCL should be considered by the cost analysis community and project managers when making decisions about programs.

Traditional bottom-up joint confidence level analysis can be cumbersome and resource intensive. This paper offers a proven top-down parametric JCL approach as a more manageable approach for cost analysts, while just as accurate as a bottom-up JCL approach based on the author's experience.

Machine learning techniques such as optimization and regression trees provide an analytical method to develop cost and schedule estimating relationships when traditional regression methods do not provide significant results.

References

Garvey, P.R., 1993, "A Family of Joint Probability Models for Cost and Schedule Uncertainties," presented at the 26th Annual Department of Defense Cost Analysis Symposium, September 1992, Williamsburg, VA.

Garvey, P.R., *Probability Methods for Cost Uncertainty Analysis: A Systems Engineering Perspective*, Marcel Dekker, New York, 2000.

Hulett, D., *Practical Schedule Risk Analysis*, Gower, Burlington, VT, 2009.

Naval Center for Cost Analysis, *Joint Agency Cost Schedule Risk and Uncertainty Handbook*, 2014, Washington, D.C.

Norman R. Augustine, *Augustine's Laws*, American Institute of Aeronautics and Astronautics, 1983, Reston, VA,1983, page 50.

Roye, K. and C.B. Smart, "Beyond Regression: Applying Machine Learning to Parametrics," presented at the 2019 ICEAA Professional Development & Training Workshop, Tampa, Florida, May 14-17, 2019.

Smart, C.B., "Maximum Likelihood Estimation for Regression of Log Normal Error," presented at the 2017 ICEAA Professional Development & Training Workshop, Portland Oregon, June 6-9, 2017.

Smart, C.B., *Solving for Project Risk Management: Understanding the Critical Role of Uncertainty in Project Management*, McGraw-Hill, New York, 2020.

Sara Jardine *is a Senior Cost Analyst for Galorath Federal with over 16 years of financial management experience. She has served a broad variety of federal agencies including the Army, Navy, OUSD AT&L, DAU, Veterans Affairs, and the Department of Homeland Security. She is skilled in Cost Analysis and course development, Project Management, Requirements Analysis, Contract Management, and Budget Management. Sara earned a MS in Project Management from George Washington University and a BS in Mathematics from the University of Michigan.*

Kimberly Roye *is a Senior Data Scientist for Galorath Federal. Starting her career as a Mathematical Statistician for the US Census Bureau, Kimberly transitioned to a career in Cost Analysis over 10 years ago. She has supported several Department of Defense hardware, software and vehicle programs, as well as NASA and the Department of Homeland Security (DHS). She is currently a lead developer of Machine Learning training for the Army and DHS. Kimberly earned a MS in Applied Statistics from Rochester Institute of Technology and a dual BS in Mathematics/Statistics from the University of Georgia.*

Dr. Christian Smart *is the Chief Data Scientist with Galorath Federal. He is author of the book Solving for Project Risk Management: Understanding the Critical Role of Uncertainty in Project Management. Dr. Smart is the VP for Professional Development with ICEAA. He regularly presents at conferences and has won several best paper awards. Dr. Smart received an Exceptional Public Service Medal from NASA in 2010 and has a PhD in Applied Mathematics.*

The International Cost Estimating and Analysis Association is a 501(c)(6) international non-profit organization dedicated to advancing, encouraging, promoting and enhancing the profession of cost estimating and analysis, through the use of parametrics and other data-driven techniques.

www.iceaaonline.com

Submissions:

Prior to writing or sending your manuscripts to us, please reference the JCAP submission guidelines found at

www.iceaaonline.com/publications/jcap-submission

Kindly send your submissions and/or any correspondence to JCAP.Editor@gmail.com

International Cost Estimating & Analysis Association

4115 Annandale Road, Suite 306 | Annandale, VA 22003

703-642-3090 | iceaa@iceaaonline.org